# A HYBRID GENETIC-NEURAL FRONT-END EXTENSION FOR ROBUST SPEECH RECOGNITION OVER TELEPHONE LINES

*Sid-Ahmed Selouani*[*], *Habib Hamam*[**], *Douglas O'Shaughnessy*[***]

[*]Université de Moncton, Campus de Shippagan, Canada
[**]Université de Moncton, Campus de Moncton, Canada
[***]INRS-Énergie-Matériaux-Télécommunications, Canada
selouani@umcs.ca, hamamh@umoncton.ca, dougo@emt.inrs.ca

## Abstract

This paper presents a hybrid technique combining the Karhonen-Loeve Transform (KLT), the Multilayer Perceptron (MLP) and Genetic Algorithms (GAs) to obtain less-variant Mel-frequency parameters. The advantages of such an approach are that the robustness can be reached without modifying the recognition system, and that neither assumption nor estimation of the noise are required. To evaluate the effectiveness of the proposed approach, an extensive set of continuous speech recognition experiments are carried out by using the NTIMIT telephone speech database. The results show that the proposed approach outperforms the baseline and conventional systems.

## 1. Introduction

Adaptation to the environment changes and artifacts remains one of the most challenging problems for the Continuous Speech Recognition (CSR) systems. The principle of CSR methods consists of building speech sound models based on large speech corpora that attempt to include common sources of variability that may occur in practice. Nevertheless, not all situations and contexts can be exhaustively covered. As speech and language technologies are being transferred to real applications, the need for greater robustness in recognition technology becomes more apparent when speech is transmitted over telephone lines, when the signal-to-noise ratio (SNR) is extremely low, and more generally, when adverse conditions and/or unseen situations are encountered. To cope with these adverse conditions and to achieve noise robustness, different approaches have been studied. Two major approaches have emerged. The first approach consists of preprocessing the corrupted speech input signal prior to the pattern matching in an attempt to enhance the SNR. The second approach attempts to establish a compensation method that modifies the pattern matching itself to account for the effects of noise. Methods in this approach include noise masking, the use of robust distance measures, and HMM decomposition. For more details see [5].

As an alternative approach, we propose a new enhancement scheme based on the combination of subspace filtering, the Multilayer Perceptron (MLP) and Genetic Algorithms (GAs) to obtain less-variant Mel-frequency parameters. The enhanced parameters are expected to be insensitive to the degradation of speech signals due to telephone-channel degradation. The main advantages of such an approach over the compensation method are that the robustness can be reached without modifying the recognition system, and without requiring assumption or estimation of the noise.

This paper is organized as follows. In section 2, we describe the basis of the signal subspace approach, namely the Karhonen-Loeve Transform (KLT) and the extension we proposed to enable the use of the technique in the Mel-frequency space. In section 3, we briefly describe the principle of MLP-based enhacement method. Then, we proceed in section 4 with the description of the evolutionary-based paradigm that we introduced to perform noise reduction. In section 5, we evaluate the hybrid MLP-KLT-GA-based front-end technique in the context of telephone speech. Finally, in section 6, we conclude and discuss our results.

## 2. Signal and Mel-frequency subspace filtering

The principle of the signal subspace techniques is based on the construction of an orthonormal set of axes. These axes point in the directions of maximum variance, thus forming a representational basis that projects on the direction of maximum variability. Applied in the context of noise reduction, these axes enable decomposing the space of the noisy signal into a signal-plus-noise subspace and a noise subspace. The enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal space. The decomposition of the space into two subspaces can be performed by using KLT (eigendecomposition). Let $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]^\mathbf{T}$ be an $N$-dimensional noisy observation vector which can be written as the sum of an additive noise distortion vector $\mathbf{w}$ and the vector of clean speech samples $\mathbf{s}$. The noise is assumed to be uncorrelated with the clean speech. Further, let $\mathbf{R_x}$, $\mathbf{R_s}$, and $\mathbf{R_w}$ be the covariance matrices from $\mathbf{x}$, $\mathbf{s}$, and $\mathbf{w}$ respectively. The eigendecomposition of $\mathbf{R_s}$ is given by $\mathbf{R_s} = \mathbf{Q}\mathbf{\Lambda_s}\mathbf{Q^T}$ where $\mathbf{\Lambda_s}$= diag$(\lambda_{s1}, \lambda_{s2}, ...., \lambda_{sN})$ is the diagonal matrix of eigenvalues given in a decreasing order. The eigenvector matrix $\mathbf{Q}$ of the clean speech covariance matrix is identical to that of the noise. Major signal subspace techniques assume the noise to be white with $\mathbf{R_w} = \sigma_\mathbf{w}^\mathbf{2}\mathbf{I}$ where $\sigma_\mathbf{w}^\mathbf{2}$ is the noise variance and $\mathbf{I}$ the identity matrix. Thus, the eigendecomposition of $\mathbf{R_x}$ is given by: $\mathbf{R_s} = \mathbf{Q}(\mathbf{\Lambda_s} + \sigma_\mathbf{w}^\mathbf{2}\mathbf{I})\mathbf{Q^T}$. The enhancement is performed by assuming that the clean speech is concentrated in an $r < N$ dimensional subspace, the so-called signal subspace, whereas the noise occupies the $N - r$ dimensional observation space. Then the noise reduction is obtained by considering only the signal subspace in the reconstruction of the enhanced signal. Mathematically it consists of finding a linear estimate of $\mathbf{s}$ given by $\hat{\mathbf{s}} = \mathbf{F}\mathbf{x} = \mathbf{F}\mathbf{s} + \mathbf{F}\mathbf{w}$ where $\mathbf{F}$ is the enhancement filter. This filter matrix $\mathbf{F}$ can be written as follows:

$\mathbf{F} = \mathbf{Q_r G_r Q_r^T}$ in which the diagonal matrix $\mathbf{G_r}$ contains the weighting factors $g_i$ with $i = 1, ..., r$, for the eigenvalues of the noisy speech. Perceptually meaningful weighting functions exist to generate $g_i$. These functions are empirically guided in order to constitute an alternative choice for $g_i$, which results in a more or less aggressive noise suppression, depending on the SNR. In [1], the linear estimation of the clean vector is performed using two perceptually meaningful weighting functions. The first function is given by :

$$g_i = \Big[\frac{\lambda_{x\,i}}{\lambda_{x\,i} + \sigma_w^2}\Big]^{\gamma}, \quad i = 1, ..., r, \tag{1}$$

where $\gamma \geq 1$.

The second function constitutes an alternative choice for $g_i$ which results in a more aggressive noise suppression:

$$g_i = \exp\Big\{\frac{-\nu\sigma_w^2}{\lambda_{x\,i}}\Big\}. \quad i = 1, ..., r, \tag{2}$$

The value of the parameter $\nu$ is to be fixed experimentally.

Instead of dealing with the speech signal, we chose to use the noisy Mel-Frequency Cepstral Coefficients (MFCC) vector $\mathbf{C}^{'}$ as well. The reason is that these parameters are suited to speech recognition due to the advantage that one can derive from them a set of parameters which are invariant to any fixed frequency-response distortion introduced by either the adverse environments or the transmission channels [6]. The main advantage of the approach proposed here is that we do not need to define weighting functions. In this approach, the filter matrix $\mathbf{F}$ can be written as follows: $\mathbf{F_{gen}} = \mathbf{Q G_{gen} Q^T}$ in which the diagonal matrix $\mathbf{G_{gen}}$ contains now weighting factors optimized using genetic operators. Optimization is reached when the Euclidian distance between the noisy and clean MFCCs is minimized. To improve the enhancement of noisy MFFCs, we introduce a preprocessing level which uses the MLP. As depicted in Figure 1, the noisy (MFCC) vectors $\mathbf{C}^{'}$ are first enhanced by MLP. Then, a KLT is performed on the output of MLP, denoted by $\hat{\mathbf{C}}$. Finally, the space of feature representation is reconstructed by using the eigenvectors weighted by the optimal factors of the $\mathbf{G_{gen}}$ matrix.

## 3. MLP-based enhancement preprocessing of the KLT

Numerous approaches were proposed in the literature to incorporate acoustic features estimated by the MLP under noisy conditions [6] [12]. The connectionist approaches offer inherent nonlinear capabilities as well as easy training from pairs of corresponding noisy and noise-free signal frames. Because the front end is very modular, the MLP estimator can be introduced at different stages in the feature processing stream. For instance, the MLP can estimate robust filterbank log-energies that will then be processed with the traditional Distrete Cosine Transform to get the unnormalized cepstral coefficients. Alternatively, we can estimate the cepstral features directly with an MLP. Yet another possibility is to estimate filterbank log-energies but to measure the feature distortion at the cepstrum level and optimize the filterbank log-energy estimator accordingly [12]. The fact that the noise and the speech signal are combined in a nonlinear way in the cepstral domain led us to

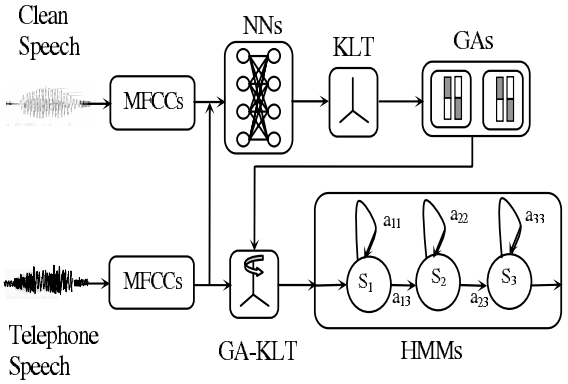


Figure 1: *The proposed MLP-KLT-GA-based CSR system.*

choose the second alternative described above. MLP can approximate the required nonlinear function to some extent [6]. Hence, the input of the MLP is the noisy MFCC vector $\mathbf{C}^{'}$, while the actual response of the network $\hat{\mathbf{C}}$ is computed during a training phase by using a convergence algorithm to update the weight vector in a manner to minimize the error between the output $\hat{\mathbf{C}}$ and the desired clean cepstrum value $\mathbf{C}$. The weights of this network are calculated during a training phase with a back-propagation training algorithm using a mean square error criterion.

The noisy 13-dimensional vector (12 MFCCs + energy) is fed to an MLP network in order to reduce the noise effects on this vector. This first-level pre-processing does not require any knowledge about the nature of the corrupting noisy signal, which permits dealing with any kind of noise. Moreover, using this enhancement technique we avoid the noise estimation process that requires a speech/non-speech pre-classification, which may be not accurate enough for low SNRs. It is worth noting that this technique is less complex than many other enhancement techniques that requires either modeling or compensating for the noise.

Once the enhanced vector is obtained, it is fed to the KLT-GA module, which represents the second enhancement level. This module refines the enhanced vector by projecting its components in the subspace generated by a genetically weighted version of the eigenvectors of the clean signal. The motivation behind the use of a second level of enhancement after using the MLP network is to compensate for the limited power of the MLP network for enhancement outside the training space [6].

## 4. Hybrid MLP-KLT-GA speech front-end

The KLT processing on the MLP-enhanced noisy vectors $\hat{\mathbf{C}}$ gives the diagonal matrix $\mathbf{G_r}$ containing the weighting factors $g_i$ with $i = 1, ..., r$. In the classical subspace filtering approaches, a key issue is to determine the rank $r$ from which the high order components (those who are supposed to contain the noise are removed). In the evolutionary-based method we propose, all components are used in the optimization process. Only the performance criterion will determine the final components that are retained to perform the reconstruction of the space of enhanced features.

The evolution process starts with the creation of a population of the weight factors, $g_i$ with $i = 1, ..., N$, which represent the individuals. The individuals evolve through many generations in a pool where genetic operators are applied [4].

Some of these individuals are selected to reproduce according to their performance. The individuals' evaluation is performed through the use of an objective function. When the fittest individual (best set of weights) is obtained, it is then use in the test phase to project the noisy data. Genetically modified MFCCs, their first and second derivatives, are finally used as enhanced features for the recognition process. As mentioned earlier, the problem of determining optimal *r* is not needed, since the GA considers the complete space dimension *N*.

### 4.1. Initialization, termination and solution representation

A solution representation is needed to describe each individual in the population. For our application, the useful representation of an individual for function optimization involves genes or variables from an alphabet of floating point numbers with values within the variables' upper and lower bounds, noted ($a_i$, $b_i$) respectively. Concerning the initialization of the pool, the ideal zero-knowledge assumption is to start with a population of completely random values of weights. These values follow an uniform distribution within the upper and lower boundaries. The evolution process is terminated when a certain number of maximum generations is reached. This number corresponds to a convergence of the objective function.

### 4.2. Selection function

A common selection method assigns a probability of selection, $P_j$, to each individual, *j*, based on its objective function value. Various methods exist to assign probabilities to individuals. In our application, the normalized geometric ranking is used [7]. This method defines $P_j$ for each individual by:

$$P_j = \frac{q(1-q)^{s-1}}{1-(1-q)^P}, \tag{3}$$

where *q* is the probability of selecting the best individual, *s* is the rank of the individual (1 is the rank of the best), and *P* is the population size.

### 4.3. Crossover

In order to avoid the extension of the exploration domain of the best solution, a simple crossover operator can be used [7]. It generates a random number $l$ from a uniform distribution and does an exchange of the genes of the parents (*X* and *Y*) on the offspring genes (*X'* and *Y'*). It can be expressed by the following equations:

$$\begin{cases} X' = lX + (1-l)Y \\ Y' = (1-l)X + lY. \end{cases} \tag{4}$$

### 4.4. Mutation

The principle of the non-uniform mutation consists of randomly selecting one component, $x_k$, of an individual $X$, and setting it equal to a non-uniform random number, $x'_k$:

$$x'_k = \begin{cases} x_k + (b_k - x_k)f(Gen) & \text{if } u_1 < 0.5 \\ x_k - (a_k + x_k)f(Gen) & \text{if } u_1 \geq 0.5, \end{cases} \tag{5}$$

where the function *f(Gen)* is given by:

$$f(Gen) = (u_2(1 - \frac{Gen}{Gen_{max}}))^t, \tag{6}$$

where $u_1$,$u_2$ are uniform random numbers in the range (0,1), *t* is a shape parameter, *Gen* is the current generation and $Gen_{max}$ is the maximum number of generations. The multi-non-uniform mutation generalizes the application of the non-uniform mutation operator to all the components of the parent *X*.

### 4.5. Objective function

The GA must search all the axes generated by the KLT of the MEL-frequency space to find the closest to the clean MFCCs. Thus, evolution is driven by a fitness function defined in terms of a distance measure between noisy MFCCs pre-processed by MLP and projected on a given individual (axis), and the clean MFCCs. The fittest individual is the axis which corresponds to the minimum of that distance. The distance function applied to cepstral (or other voice representations) refers to *spectral distortion measures* and represents the cost in a classification system of speech frames. For two vectors $\mathbf{C}$ and $\hat{\mathbf{C}}$ representing two frames, each with *N* components, the geometric distance is defined as:

$$d(\mathbf{C}, \hat{\mathbf{C}}) = (\sum_{k=1}^{N} (\mathbf{C_k} - \hat{\mathbf{C}}_\mathbf{k})^l)^{1/l}. \tag{7}$$

For simplicity, the Euclidian distance is considered ($l = 2$), which turned out to be a valuable measure for both clean and noisy speech. Note that $-d(\mathbf{C}, \hat{\mathbf{C}})$ is used as a distance measure because the evaluation function must be maximized.

## 5. Experiments and Results

Extensive experimental studies were carried out to characterize the impairment induced by telephone networks [3]. When speech is recorded through telephone lines, a reduction in the analysis bandwidth yields a higher recognition error, particularly when the system is trained with high-quality speech and tested using simulated telephone speech [9].

In our experiments, the training set composed of the $dr1$ and $dr2$ subdirectories of the TIMIT database, described in [2], was used to train a set of clean speech models. The speech recognition system used the $dr1$ subdirectory of NTIMIT as test set [2]. HTK the HMM-based speech recognition system described in [11] has been used throughout all experiments. We compared three systems: the KLT-based system as detailed in [10], the new MLP-KLT-GA-based CSR system and the baseline HMM-based system which uses a MFCC+first and second derivatives front-end denoted: MFCC_D_A. The architecture of the MLP network consists of three layers. The input layer consists of 13 neurons, while the hidden layer and the output layer consists of 26 and 13 neurons, respectively. The input to the network is the noisy 12-dimensional MFCC vector in addition to the energy. The weights of this network are calculated during a training phase with a back-propagation algorithm with a learning rate equal to 0.25 and a momentum coefficient equal to 0.09. The obtained weight values are then used during the recognition process to reduce the noise in the enhanced obtained vector that is incorporated into the KLT-GA module. To control the run behaviour of a genetic algorithm, a number of parameter values must be defined. The initial population is composed of 250 individuals and was created by duplicating the elements of the weighting matrix. The genetic algorithm was halted after 500 generations. The percentages of crossover rate and mutation rate are fixed respectively at 28% and 4%. The number of

| | $\%\epsilon_{Sub}$ | $\%\epsilon_{Del}$ | $\%\epsilon_{Ins}$ | $\%C_{Wrd}$ |
|---|---|---|---|---|
| MFCC_D_A (1) | 82.71 | 4.27 | 33.44 | 13.02 |
| KLT-(1) | 77.05 | 5.11 | 30.04 | 17.84 |
| MLP-KLT-GA-(1) | 52.15 | 5.07 | 21.36 | **43.22** |

[a] $\%C_{Wrd}$ using 1-mixture triphone models.

| | $\%\epsilon_{Sub}$ | $\%\epsilon_{Del}$ | $\%\epsilon_{Ins}$ | $\%C_{Wrd}$ |
|---|---|---|---|---|
| MFCC_D_A (1) | 81.25 | 3.44 | 38.44 | 15.31 |
| KLT-(1) | 78.11 | 3.81 | 48.89 | 18.08 |
| MLP-KLT-GA-(1) | 49.78 | 3.68 | 49.40 | **46.48** |

[b] $\%C_{Wrd}$ using 2-mixture triphone models.

| | $\%\epsilon_{Sub}$ | $\%\epsilon_{Del}$ | $\%\epsilon_{Ins}$ | $\%C_{Wrd}$ |
|---|---|---|---|---|
| MFCC_D_A (1) | 78.85 | 3.75 | 38.23 | 17.40 |
| KLT-(1) | 76.27 | 4.88 | 39.54 | 18.85 |
| MLP-KLT-GA-(1) | 50.95 | 3.58 | 22.98 | **49.10** |

[c] $\%C_{Wrd}$ using 4-mixture triphone models.

| | $\%\epsilon_{Sub}$ | $\%\epsilon_{Del}$ | $\%\epsilon_{Ins}$ | $\%C_{Wrd}$ |
|---|---|---|---|---|
| MFCC_D_A (1) | 78.02 | 3.96 | 40.83 | 18.02 |
| KLT-(1) | 77.36 | 5.37 | 34.62 | 17.32 |
| MLP-KLT-GA-(1) | 47.85 | 5.86 | 25.39 | **50.48** |

[d] $\%C_{Wrd}$ using 8-mixture triphone models.

Table 1: *Percentages of word recognition rate ($\%C_{Wrd}$), insertion rate ($\%\epsilon_{Ins}$), deletion rate ($\%\epsilon_{Del}$), and substitution rate ($\%\epsilon_{Sub}$) of the MFCC_D_A_ (denoted (1)), KLT-MFCC_D_A, and MLP-KLT-GA-MFCC_D_A CSR systems using (a) 1-mixture, (b) 2-mixture, (c) 4-mixture and (d) 8-mixture* **tri-phone** *models. (Best rates are highlighted in boldface.)*

total runs was fixed at 70. After the GA processing, the MFCCs static vectors are then expanded to produce a 39-dimensional (static+dynamic) vector upon which the hidden Markov models (HMMs), that model the speech subword units, were trained.

We found through experiments that using the MLP-KLT-GA as a pre-processing approach to enhance the MFCCs that were used for recognition with $N$-mixture Gaussian HMMs for $N$=1, 2, 4 and 8, using tri-phone models, leads to an important improvement in the accuracy of the word recognition rate. A correct rate of 50.48% is reached by the MLP-KLT-GA-MFCC_D_A-based CSR system when the baseline and the KLT-baseline systems achieve 18.02% and 17.32% respectively. This represents an improvement of more than 32% comparatively to the baseline system when the 8-mixture tri-phone model is used. Expanding to more than 8 mixtures did not improve the performance. The results in Table 1 show also that substitution and insertion errors are considerably reduced when the hybrid neural-evolutionary approach is included, leading to more effectiveness to the CSR system.

## 6. Conclusion

In this paper, a hybrid genetic-neural front-end was proposed to improve speech recognition over telephone lines. It is based on an MLP-KLT-GA hybrid enhancement scheme which aims to obtain less-variant MFCC parameters under telephone-channel degradation. Experiments show that the use of the proposed robust front-end processing increases the recognition rate by 32% when *dr1* and *dr2* TIMIT directories are used for the training and *dr1* directory of NTIMIT for the test. This indicates that both subspace filtering and GA-based optimization gained from the use of MLP as pre-processing. It is worthy of noting that the neural-evolutionary-based technique is less complex than many other enhancement techniques, which need to either model or compensate for the noise. For further work, many other directions remain open. Present goals include the improvement of the objective function in order to perform the online adaptation of the HMM-based CSR system when it faces new and unseen contexts and environments.

## 7. References


[1] Ephraim Y., and Van Trees H. L., "A signal subspace approach for Speech Enhancement", IEEE Transactions on Speech and Audio Processing, 3(4), pp. 251–266, 1995.

[2] Fisher W. M., Dodington G. R., and Goudie-Marshall K. M., "The DARPA Speech Recognition Research Database: Specification and Status", Proc. DARPA Workshop on Speech Recognition, pp. 93–99, 1986.

[3] Gaylor W. D., "Telephone voice transmission. standards and measurements", Prentice Hall, Englewood Clifs, N.J. 1989.

[4] Goldberg D. E., "Genetic algorithms in search, optimization and machine learning", Addison-Wesley Publishing, 1989.

[5] Gong Y., "Speech Recognition in Noisy Environments: A survey", Speech Communication, 16, pp. 261-291, 1995.

[6] Haverinen H., Salmela P., Hakkinen J., Lehtokangas M., and Saarinen J., "MLP Network for Enhancement of Noisy MFCC Vectors", Proc. Eurospeech, pp. 2371-2374, 1999.

[7] Houk C. R., Joines J. A. and Kay M. G., "A Genetic Algorithm for function optimization: a matlab implementation", North Carolina University-NCSU-IE, TR 95-09, 1995.

[8] Jankowski C., Kalyanswamy A., Basson S. and Spitz J., "NTIMIT: A phonetically balanced continuous speech, telephone bandwidth speech database", Proc. IEEE-ICASSP, Vol.1, pp.109–112, 1990.

[9] Moreno P. J., and Stern R., "Sources of degradation of speech recognition in the telephone network", Proc. IEEE-ICASSP, Vol.1, pp. 109-112, 1994.

[10] Selouani S.-A., and O'Shaughnessy D., "Robustness of speech recognition using genetic algorithms and a Mel-cepstral subspace approach", Proc. IEEE-ICASSP, Vol.I, pp. 201–204, 2004.

[11] Speech Group, Cambridge University , "The HTK Book (Version 3.4)", Cambridge University Group, 2006.

[12] Weintraub M., and Beaufays F., "Increased Robustness of Noisy Speech Features Using Neural Networks", Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, May 1999.